[Title Page]

Research on Multi-factor Forest Fire Prediction Model Using Machine Learning Method in China

Yudong Li,[1] Zhongke Feng,[1] Ziyu Zhao,[1]Shilin Chen,[1]Hanyue Zhang,[1]

1 Precision Forestry Key Laboratory of Beijing, Beijing Forestry University, Beijing 100083, China.

E-mail address ：

Yudong Li：lyd85842@163.com**;** **Zhongke Feng：**zhongkefeng@bjfu.edu.cn

Ziyu Zhao：zhaozy0315@126.com**;** **Shilin Chen:** chenshilin@bjfu.edu.cn

Hanyue Zhang：hanyue.zhang@foxmail.com

Correspondence should be addressed to Zhongke Feng: zhongkefeng@bjfu.edu.cn

Permanent address: Beijing Forestry University

[Manuscript]

Research on Multi-factor Forest Fire Prediction Model Using Machine Learning Method in China

Yudong Li,[1] Zhongke Feng,[1] Ziyu Zhao,[1]Shilin Chen,[1]Hanyue Zhang,[1]

1 Precision Forestry Key Laboratory of Beijing, Beijing Forestry University, Beijing 100083, China.

Correspondence should be addressed to Zhongke Feng: zhongkefeng@bjfu.edu.cn

Abstract

Forest fires can cause serious harm in many ways. Studying the scientific prediction of forest fires is an

Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

important basis for preventing such fires. At present, there is little research on the prediction of long time series forest fires in China. Choosing a suitable forest fire prediction model is of great importance to China's forest fire prevention and control work. Based on data on fire hotspots, meteorology, terrain, vegetation, infrastructure, and socio-economics collected from 2003 to 2016, we used a random forest model as a feature-selection method to determine 13 major drivers of forest fires in China (such as temperature, terrain etc.). The forest fire prediction models developed in this study are based on four machine-learning algorithms: an artificial neural network, a radial basis function network, a support-vector machine, and a random forest. The models were evaluated using the five performance indicators of accuracy, precision, recall, f1 value, and area-under-the-curve value. We used the optimal model to obtain the probability of forest fire occurrence in various provinces in China and create a spatial distribution map of the areas with high incidences of forest fires. The results show that the prediction accuracy of the four forest fire prediction models is between 75.8% and 89.2%, and the area-under-the-curve value is between 0.840 and 0.960. The random forest model has the highest accuracy (89.2%) and area-under-the-curve value (0.96). It is used as the optimal model to predict the probability of forest fire occurrence in China. The prediction results indicate that the areas with high incidences of forest fires are mainly concentrated in northeastern China (Heilongjiang Province and northern Inner Mongolia Autonomous Region), southeastern China (including Fujian Province and Jiangxi Province) etc. In those areas at high risk of forest fires, the management departments can improve the forest fire prevention and control by establishing watch towers and using other monitoring equipment. This study not only helps in understanding the main drivers of forest fires in China, but it also provides a reference for the selection of high-precision forest fire prediction models and provides a scientific basis for China's forest fire prevention and control work.

Keywords: forest fire occurrence in China; feature selection; forest fire driving factors; machine learning; prediction model; forest fire prevention and control

1. Introduction

Forest fires are one of the most dangerous natural disasters. They have become the focus of worldwide attention due to their rapid spread, their low controllability, and the hazards they pose[1-2]. Forest fires have varying degrees of impact on human health and safety, the ecological environment and resources, and society and the economy[3]. Forest fire prevention has therefore become a key research topic in the fields of forestry and ecology[4-6].

The most effective way to control forest fires is to detect them quickly. Detection is usually divided into three categories: satellite monitoring, smoke detection, and local perception (such as data analysis). Satellite monitoring is expensive, involves delays, and is not fully applicable to all locations[7] . Smoke detection also requires expensive equipment and maintenance work. In contrast, data for forest fire analysis are easy to obtain, and data analysis is less expensive than the first two methods[8]. By establishing a forest fire prediction model, we can predict the probability of the occurrence of a forest fire and then strictly manage the area where the fire is likely to occur. This can directly reduce the occurrence of forest fires and the potential casualties and economic losses[9-10]. This method is therefore of great significance in forest fire prediction and prevention[11].

Much research has been conducted on forest fire prediction models. Logistic regression models are the most commonly used. They have the advantage of solving the classification problem[12-16]. In recent years, geographically weighted regression models have also been used. Wang[17]used a geographically weighted regression model to predict regional fires in Gansu, China (2017). Guo[18] used a geographically weighted logistic regression model to determine the relationship between human-made fires and the potential drivers of forest fires in northern China(2016). Such a method can provide a reasonable explanation for spatial heterogeneity, but the regression analyses can only be performed on continuous variables; the method lacks analysis of categorical variables. Liu [19] used an exponential equation to predict the number of forest fires in China, but this model analyzed only meteorological factors(2017). Similarly, many researchers have used generalized linear regression models for forest fire prediction. Miao et al.[20] used the zero-inflated Poisson model to predict the frequency of forest fires in Japan in 2000(2008). Mandallaz et al. [21] used the Poisson model to predict forest fires in France, Italy, etc. (1997). Guo et al. [22] used ordinary least squares regression, zero-inflated negative binomial regression and the zero-inflated negative binomial model to predict the number of forest fires in the Greater Xing'an Mountains area of Heilongjiang Province, China, and demonstrated that the zero-inflated negative binomial model has the best performance(2010).

The development of artificial intelligence has led researchers to focus on building a forest fire prediction model using machine-learning algorithms[23-31]. Artificial neural networks consist of neurons with adjustable connection weights. Compared with traditional multiple linear regression models or parametric regression models, neural networks have better self-organization and self-learning capabilities. They have been widely used in forest fire prediction[32-34]. For example, Maeda et al.[35] used artificial neural networks and multi-temporal images from MODIS/Terra-Aqua sensors to detect areas at high risk of forest fires in the Amazon region of Brazil(2009). The results showed that the error is less than 1, and the predictions are accurate. Sakr et al. [36] predicted the occurrence of forest fires in developing countries through two meteorological factors using artificial neural networks (2011).

A radial basis function (RBF) neural network is a three-layer neural network. It is a special case of back propagation neural network. At present, little research has used RBF neural networks for forest fire prediction. Samaher[25] used an RBF neural network to predict the forest fire risk in Portuguese natural parks (2018). The final root mean square error was 54.2.

Support-vector machines (SVM) are most suitable for binary classification of data in the form of supervised learning. They apply the principle of structural risk minimization and have good learning ability. In recent years, researchers have begun to use SVMs to predict forest fires[8][37-40]. Samaher [25] used five different soft computing (SC) technologies, including an SVM algorithm, to predict areas at risk of forest fires (2018). He determined that the SVM algorithm provides more accurate prediction than the other four algorithms. Cortez et al. [8] used five different Data Mining(DM) algorithms to predict the area at risk of forest fires in the northeastern region of Portugal (2007). Their results showed that the root mean square error was 64.7. Based on Cortez's research, Zhiqing et al.[7]used the semi-definite programming model to select the optimal kernel function of the SVM to establish an SVM model for forest fire prediction (2012). The mean square error was 1.76, and the model effect was good.

The random forest (RF) algorithm is a well-known integrated learning algorithm. It can provide higher accuracy than other algorithms. At present, the use of RFs to predict forest fires is relatively established[41-43]. Liang et al. [44] used an RF model to predict the occurrence of forest fires in Fujian Province, China, with an accuracy rate of 85% (2016). Pourtaghi et al. [45] used an RF algorithm to study the sensitivity of forest fires in Golestan Province, Iran. Their results showed that the model achieves the desired accuracy (2016).

Most of the current research focuses on certain areas, and there are few studies on the prediction and analysis of long time series in China. Most such studies concentrate on the temporal and spatial changes and influencing factors of forest fires in specific years[46-50]. The results of previous research are therefore localized and limited, and there is a lack of research into the most suitable and high-precision forest fire prediction model on the national scale.

In this study, we selected a variety of forest fire driving factors to build four prediction models based on machine-learning algorithms. The models were evaluated using data on Chinese forest fires from 2003 to 2016. The study has three objectives: (1) identify the main forest fire driving factors and their impact in China; (2) select the most suitable model for forest fire prediction in China by creating four models and comparing and analyzing the fitting results; and (3) use the model that offers the most accurate predictions to create a probability map of forest fires in China and put forward recommendations for forest fire prevention.

2. Materials and Methods

2.1 Study Area and Data Resources

Located in east Asia on the west coast of the Pacific Ocean, China's territory is vast, with a total land area of about 9.6 million square kilometers. The topography is high in the west, with vast mountains and plateaus, and low in the east. The distance between the east and the west of the country is about 5,000 kilometers; the coastline of the mainland is more than 18,000 kilometers in length; and the temperature and precipitation are diverse, forming a variety of climates[51]. The distribution of forest resources in China is uneven, being mainly distributed in the northeast, south, and southwest regions. The forested area is 220 million hectares, and the forest coverage rate is 22.96%.

The research data were divided into six parts: fire ignition data, meteorological data, terrain data, vegetation data, infrastructure data, and socio-economic data. The fire point data were derived from NASA's *Global Fire Atlas with Characteristics of Individual Fires, 2003–2016* (https://daac.ornl.gov/). The *Global Fire Atlas* is a global dataset that tracks the daily dynamics of single fires. For each individual fire, the dataset provides information about the fire's timing and location, scale, perimeter, duration, speed, and direction of spread. These individual fire characteristics are based on the *Global Fire Atlas* algorithms and estimated combustion day information from a 500-meter resolution product of the 6 MCD64A1 combustion zone product of the Medium Resolution Imaging Spectroradiometer (MODIS) collection.

This study used the fire point data for forest land in China from 2003 to 2016. The final number of fire point data are 32746 (except Taiwan). The meteorological data are derived from the 14-day daily value dataset of the China Meteorological Data Network (http://data.cma.cn/dataService/). The dataset includes eight elements, such as the barometric pressure, temperature, relative humidity, and precipitation of the station. Digital elevation model data were obtained through the Geospatial Data Cloud website (http://www.gscloud.cn/). Vegetation data were represented by the normalized difference vegetation index (NDVI), and the spatial distribution dataset of China's Quarterly Vegetation Index comes from the Resource and Environment Data Cloud Platform (http://www.resdc.cn/). The basic geographic data were taken from the "National Basic Geographic Database of 1:1 Million" on the website of the National Geographic Information Resources Directory System. The data include the locations of railways, highways, water systems, and residential areas. The socio-economic data include population density and GDP per capita, and the grid data of the spatial distribution of population and GDP were obtained from the Resource and Environment Data Cloud Platform. Figure 1 shows the map of the study area.

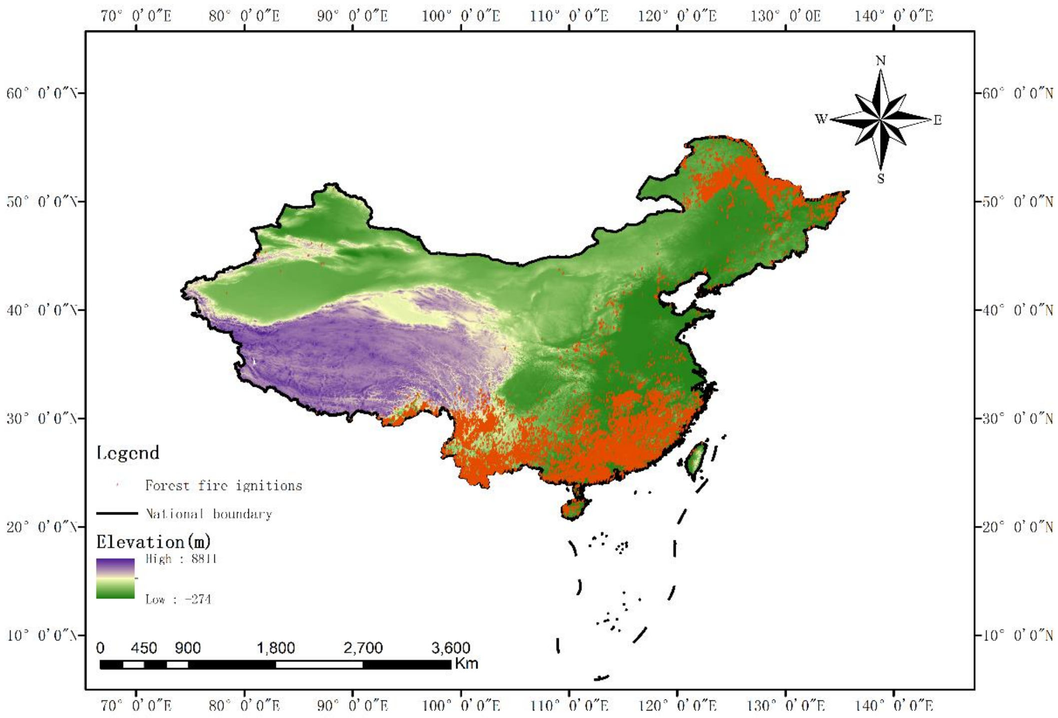


Fig.1 Map of the study area

2.2 Data Preprocessing

2.2.1 Variable Handling

The dependent variable is a binary variable (i.e., whether a forest fire occurs), and so we used ArcGIS to create a certain percentage of random points (non-fire points) and assigned 1 to fire points and 0 to non-fire points[52]. To ensure that the data were not over-dispersed, random points were selected according to experience in a ratio of 1:1[53], and, in principle, randomness in space and time should be followed[54]. We used the ArcGIS 10.4 software to create random points and then used the 2015 National Land Use data as a basis to exclude random points that fell in bodies of water or urban land. We obtained a total of 65,492 fire points and random points.

For the meteorological data, we first used the ArcGIS 10.4 software to match the sample points with the nearest meteorological station through the Thiessen polygon method. We then extracted the corresponding sample point weather data and used a SQL Server database to match the daily weather data. For the terrain data, we used the spatial analysis tool in the ArcGIS 10.4 software to extract the slope and aspect of the obtained digital elevation model data. Seasonal climatic differences have an impact on vegetation status, and so we divided the year into spring (March, April, May), summer (June, July, August), autumn (September, October, November), and winter (December, January, February) [55-56]. We used the extraction and analysis tools of the ArcGIS software to extract NDVI data for the sample points on an annual and quarterly basis.

Similarly, from the infrastructure data and socio-economic data, we extracted the information corresponding to the sample points. We set the aspect and special festivals as categorical variables, and the others as continuous variables. Table 1 shows the classification of aspect [57]. During certain traditional festivals in China, people burn paper to commemorate their loved ones, which raises the probability of a forest fire. We included as special festivals (value 1) the following dates: Chinese New Year's Eve, the first day of the first lunar month, the second day of the first lunar month, the fifteenth day of the first lunar month, and Qingming Festival and Zhongyuan Festival (July 15th of the lunar calendar). Non-special festivals were set to 0.

Table 1: Descriptions of aspect classifications

Aspect	Azimuth (degree)	Classification
Gentle slope	-1	0
Shady slope	0~67.5, 337.5~360	1

Semi-shady slope	67.5~112.5, 292.5~337.5	2
Sunny slope	157.5~247.5	3
Semi-sunny slope	112.5~157.5, 247.5~292.5	4

After processing, we obtained 20 independent variables and their possible values (see Table 2). Finally, we performed data cleaning on the sample points and the various types of data extracted to remove abnormal samples from the original dataset (including some samples with missing data and samples with observations that were significantly outside the normal range).

Table 2: Descriptions of independent variables

Category	Independent Variable	Symbol	Variable Type
Location	Longitude (°)	*Lon*	Continuous Variable
	Latitude (°)	*Lat*	Continuous Variable
Terrain	Altitude (m)	*Alt*	Continuous Variable
	Slope (°)	*Slo*	Continuous Variable
	Aspect	*Asp*	Categorical Variable
Meteorology	Average surface temperature (°C)	*Avst*	Continuous Variable
	Daily maximum surface temperature (°C)	*Mast*	Continuous Variable
	Cumulative precipitation at 20–20 (mm)	*Pre*	Continuous Variable
	Average relative humidity (%)	*Arh*	Continuous Variable
	Hours of sunshine (h)	*Suh*	Continuous Variable
	Average temperature (°C)	*Ate*	Continuous Variable
	Daily maximum temperature (°C)	*Mate*	Continuous Variable
	Average wind speed (m/s)	*Aws*	Continuous Variable

	Maximum wind speed (m/s)	*Mws*	Continuous Variable
Infrastructure	Distance from fire point to highway (m)	*Hig*	Continuous Variable
	Closest distance from fire point to residential area (m)	*Set*	Continuous Variable
Social humanity	Population	*Pop*	Continuous Variable
	GDP	*GDP*	Continuous Variable
	Special festival	*Sfe*	Categorical Variable
Vegetation	NDVI	*NDVI*	Continuous Variable

2.2.2 Data Normalization

Given the different dimensions and magnitudes of the factors above, the data were normalized to eliminate the variation in dimensions, avoid large differences in the magnitudes of the input and output data, and balance the contributions of various factors. All the data were converted to between 0 and 1. Table 3 shows the normalized formulas and specific interpretations of the independent variables.

Table 3: Normalized formulas and explanations

No.	Formula	Explanation	Variables using this formula
(1)	$x_i^* = \dfrac{x_i - x_{min}}{x_{max} - x_{min}}$	x_i and x_i^* are the values before and after data normalization; x_{max} and x_{min} are the maximum and minimum values of the full sample data.	*Lon, Lat, Alt, Avst, Mast, Pre, Suh, Ate, Mate, Aws, Mws, Hig, Set, Pop, GDP*
(2)	$x_\alpha = \sin\alpha$	α is the slope value	*Slo*
(3)	$x_\gamma = \dfrac{\gamma}{100}$	γ is the humidity value	*Arh*

2.3 Research Method

2.3.1 Artificial Neural Networks

Artificial neural networks (ANN) have become widely used in feedforward networks due to their clear structure, fast operation, easy implementation, and abilities for self-learning and adaption to the environment [24][58-59]. ANNs consist of three parts: an input layer, an output layer, and a hidden layer. The hidden layer may be a topological structure of one or more layers, as shown in Figure 2. The input layer does not perform any calculations. It is used to receive data; that is, to transfer data to the adjacent hidden layer with different weights. The hidden layer processes the data through a nonlinear activation function and then passes it to the output layer. The final result is obtained from the output layer [60]. The mathematical principle is as follows:

$$\begin{cases} \mathrm{h}^{(1)} = \varphi^{(1)}(\sum_{i=1}^{n} x_i \cdot {\omega_j}^{(1)} + b^{(1)} \\ y = \varphi^{(2)}(\sum_{j=1}^{n} {h_i}^{(1)} \cdot {\omega_j}^{(2)} + b^{(2)} \end{cases} \quad (4)$$

In the formula, input layer $x \in R^m$, hidden layer output $\mathrm{h} \in R^n$, output layer $y \in R^K$, input layer to hidden layer weight connection matrix $\omega^{(1)} \in R^{m \times n}$, the weight connection bias from the input layer to the hidden layer $b^{(1)} \in R^n$, the weight connection matrix and the bias from the hidden layer to the output layer are $\omega^{(2)} \in R^{n \times K}$ and $b^{(2)} \in R^{n \times K}$.

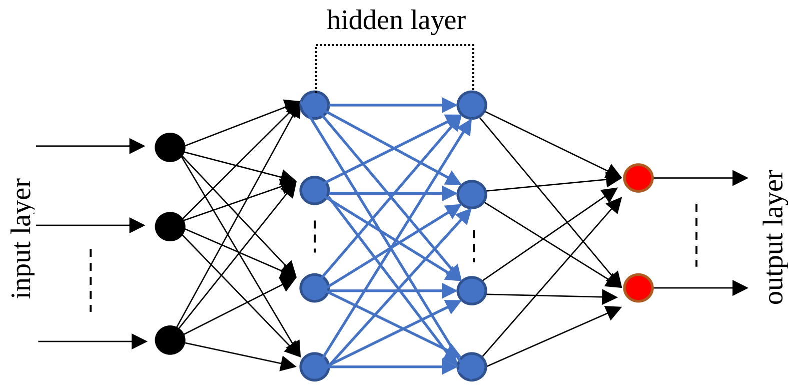


Fig. 2 Diagram of the structure of an ANN

2.3.2 Radial Basis Function Neural Network

The radial basis function (RBF) neural network structure is a feedforward structure with an input layer, a single hidden layer, and an output layer [61]. Its advantages are concise training and fast learning convergence speed, which can approximate any nonlinear function. It has been widely used in time-series forecasting, nonlinear control systems, and the graphics-processing field. The basic idea of an RBF network is as follows. The RBF is used as the "base" of the hidden unit to form the hidden layer space. The hidden layer transforms the input vector and transforms the low-dimensional pattern input data into the high-dimensional space. The result is that the data are linearly separable in the high-dimensional space. The output of the RBF neural network is:

$$y_i = \sum_{i=1}^{h} \omega_{ij} \exp\left(-\frac{1}{2\sigma^2}\left\|x_p - c_i\right\|^2\right) \quad j = 1,2,\cdots,n \tag{5}$$

where $x_p = (x_1{}^p, x_2{}^p, \cdots, x_m{}^p)^T$ is the p-th input sample (p = 1,2,3,...,P), P is the total number of samples, c_i is the center of the hidden layer node of the network, ω_{ij} is the connection weight from the hidden layer to the output layer, i = 1,2,3,...,h is the number of hidden layer nodes, y_i is the actual output of the j-th output node of the network corresponding to the input sample [62].

2.3.3 Support-Vector Machines

Support-vector machines (SVM) are mainly used for pattern classification and nonlinear regression. They are general learning algorithms based on the principle of structural risk minimization. The core idea of SVMs is to establish a classification hyperplane as a decision surface to maximize the isolation edge between the positive and negative examples, thereby providing a high generalization performance[63]. SVMs can improve the ability to transform data from high-dimensional spaces by flexibly using kernel functions when dealing with various nonlinear problems. Taking a two-class SVM as an example, given a training set $T = \{(x_1, y_1), \cdots (x_l, y_l)\} \in (X \times Y)^l$, where $x_i \in X = R^n, y_i \in \{1, -1\}(i = 1,2, \cdots l)$, x_i is the feature vector. The penalty parameter C and the kernel function $K(x, x')$ are first selected, and the optimization problem is then constructed and solved as follows [62]:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{j} \sum_{j=1}^{l} y_i y_j a_i a_j K(x, x') - \sum_{j=1}^{l} \alpha_j \tag{6}$$

$$s.t.\ \sum_{i=1}^{l} y_i \alpha_i = 0,\ 0 \leq \alpha_i \leq C, i = 1, \cdots, l \tag{7}$$

The optimal solution is then obtained: $\alpha^* = ({\alpha_1}^*, \cdots, {\alpha_l}^*)^T$. A positive component of α^* :$0 \leq {\alpha_j}^* \leq C$ is then selected, and the threshold is calculated as follows:

$$b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^{\cdot} K(x_i - x_j) \tag{8}$$

Finally, the decision function is constructed:

$$f(x) = sgn(\sum_{i=1}^{l} {\alpha_i}^* y_i K(x, x_i) + b^* \tag{9}$$

2.3.4 Random Forest

A random forest is a highly flexible machine-learning algorithm with wide application prospects. In essence, an RF is a classifier consisting of multiple decision trees formed by random methods. These trees are not related, hence its alternative name: “random decision tree.” When the test data enter the RF, each decision tree is classified, and the category with the most classification results among all the decision trees is taken as the final result. The RF algorithm has the following advantages: it evaluates the importance of each feature in classification problems, it can process input samples with high-dimensional features, and it does not require a reduction in dimensionality. The method is as follows.

Let *N* be the number of attributes of the sample. *n* is an integer greater than 0 and less than *N*. First, the bootstrap method is used for resampling, randomly generating *M* training sets S1, S2, ...SM. The decision tree A1, A2, ...AM corresponding to each training set is then generated. Before selecting the attribute in each non-leaf node, *n* attributes are randomly selected from the *N* attributes as the split attribute set of the current node, and the node is split in the best split mode among the *n* attributes. Each tree grows intact without pruning. For the test set sample *X*, each decision tree is used to test and obtain the corresponding categories C1(X), C2(X), ...CM(X). Finally, the voting method is adopted, and the category with the most output among the *M* decision trees is regarded as the category to which the test set sample *X* belongs [62].

2.3.5 Model Performance Evaluation

In this study, we used five performance indicators: accuracy, precision, recall, f1 value, and area-under-the-curve (AUC) value to evaluate the performance of the models. Descriptions of the five indicators are given below.

1. Accuracy: the proportion of the number of samples (TP and TN) that are correctly predicted to the

total number of samples. The formula is as follows:

$$P = \frac{TP+TN}{TP+FP+TN+FN} \tag{10}$$

2. Precision: characterizes the classification effect of the classifier, which is the correct frequency value predicted in the instance of the positive sample:

$$T = \frac{TP}{TP+FP} \tag{11}$$

3. Recall: characterizes the recall effect of a certain class. It is the correct frequency of prediction in the instance of the label as the positive sample:

$$R = \frac{TP}{TP+FN} \tag{12}$$

4. f1 value: the value used to measure precision and recall. It is the harmonic mean of these two values:

$$\mathrm{f1} = \frac{(1+a^2)PR}{a^2(P+R)} \tag{13}$$

5. A ROC (receiver operating characteristic) curve is a method to judge the prediction effect of the model[63]. The prediction accuracy of the model is judged by the value of the area under the curve (AUC). The AUC ranges from 0.5 to 1. The larger the value, the closer the fit of the model.

Note: TP, FN, FP, TN in the formulas are the labels of the confusion matrix form of the output result. The form is shown in Table 4:

Table 4: Confusion matrix form

Prediction (column) / label (row)	Positive sample	Negative sample

Positive sample	TP	FN
Negative sample	FP	TN

3. Results

In this study, we used the MATLAB (MathWorks, USA, MATLAB 2019a) [58] and RStudio (JJ Allaire, RStudio-1.2.5042/R 3.6.3) programming languages to implement the algorithms. We used MATLAB to build the SVM, ANN, and RBFN models and used RStudio to build the RF models.

To evaluate feature factors and model performance issues, the dataset was divided into two parts by randomly selecting 70% of the preprocessed sample data as the training set and 30% as the test set [59].

3.1 Feature Selection

We used the RF algorithm to perform feature selection on all variables after preprocessing, and we selected the subset of features that have the greatest impact on the dependent variable for the next model-building process. We divided the full sample according to the above-mentioned proportions (70% of the training set and 30% of the test set), and we obtained five training samples after repeating the process five times. Then we used the varSelRF package in the R language to perform feature variable selection calculations on the five training samples to obtain the variable subsets of the five intermediate models. Finally, we chose the variables that appeared 3 times or more in the five variable subsets as the main forest fire driving factors to enter the model fitting process. Table 5 shows the results of feature selection.

Table 5: Results of variable selection based on RF

No.	Variable	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Frequency

1	Lat	+	+	+	+	+	5
2	Lon	+	+	+	+	+	5
3	Avst	+	+	+	+	+	5
4	Mast	+	+		+	+	4
5	Pre	+	+		+	+	4
6	Arh	+	+	+	+	+	5
7	Suh	+	+	+	+	+	5
8	Ate	+	+	+	+	+	5
9	Mate	+	+	+	+	+	5
10	Aws						0
11	Mws						0
12	Alt	+	+	+	+	+	5
13	Slo						0
14	Asp						0
15	Set						0
16	Hig						0
17	GDP	+	+		+	+	5
18	Pop	+	+	+	+	+	5
19	NDVI	+	+	+	+	+	5
20	Sfe						0

The results show that the main influencing variables are longitude, latitude, average surface temperature, daily maximum surface temperature, accumulated precipitation, average relative humidity, sunshine hours, average temperature, daily maximum temperature, altitude, population, GDP, and NDVI. These variables performed subsequent model fitting. Then the mean decrease accuracy obtained by the RF algorithm was used to evaluate the importance of the variable. The larger the value, the greater the importance of the variable. Figure 3 shows the importance of each variable in five random training samples and 20 feature subsets in the full sample.

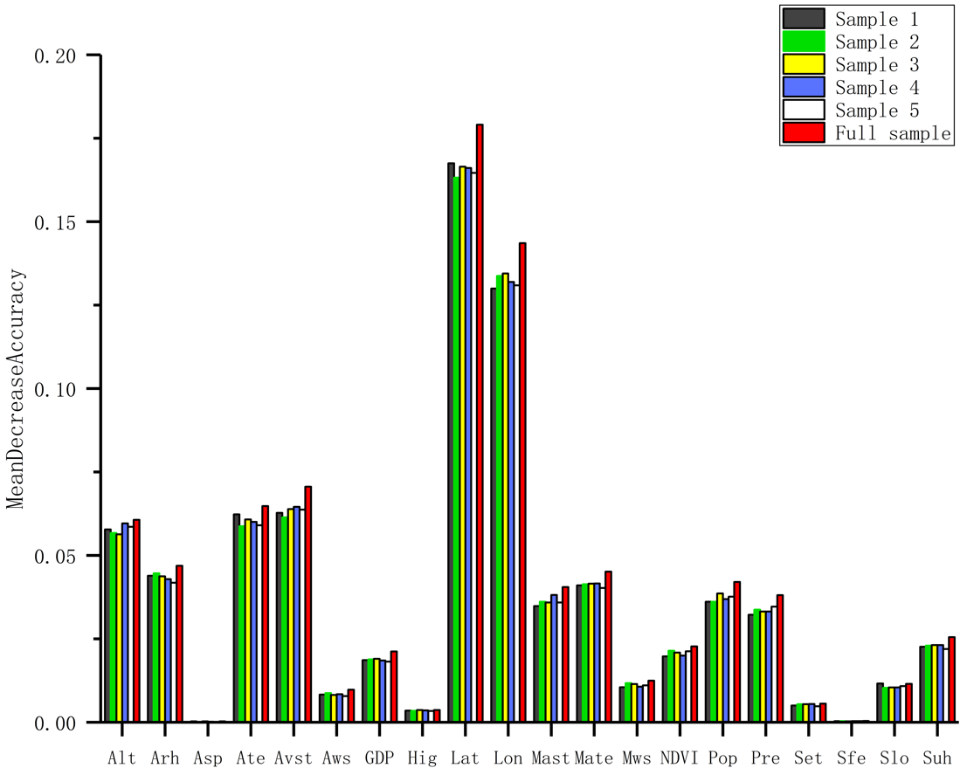


Fig. 3 Feature subset importance

It can be seen from Figure 3 that the longitude and latitude have the greatest influence on the occurrence of forest fires. Thus, the location factor has the most influence on the occurrence of forest fires. In addition, altitude also affects the occurrence of forest fires. The second greatest influence is the temperature factor (average surface temperature and average temperature), reflecting the fact that high temperatures can cause fires. Meteorological factors, including rainfall, sunshine hours, and average relative humidity, can cause forest fires to varying degrees. Human activities (GDP and population) and vegetation coverage also have an influence on the occurrence of forest fires but less so than other factors such as weather and location. The variables not selected by the RF algorithm include average wind speed, maximum wind speed, aspect, slope, the closest distance from the fire point to the highway, the closest distance from the fire point to a residential area, and special festivals. The results indicate that these seven variables have little influence on the occurrence of forest fires during the data analysis.

3.2 Model Fitting Results

3.2.1 Artificial Neural Network

The input layer of the ANN consists of 13 neurons after feature selection: Lat, Lon, Avst, Mast, Pre, Arh, Suh, Ate, Mate, Alt, GDP, Pop, and NDVI. The output layer contains two units (1 or 0). We used gradient descent to optimize the algorithm. Finally, we used a single hidden layer containing five units. The comparison between the predictive value and the actual value in the test dataset is shown in Figure 4. Note: Due to the large sample size, only a part of the sample comparison chart is displayed. This is also the case for the following comparison charts.

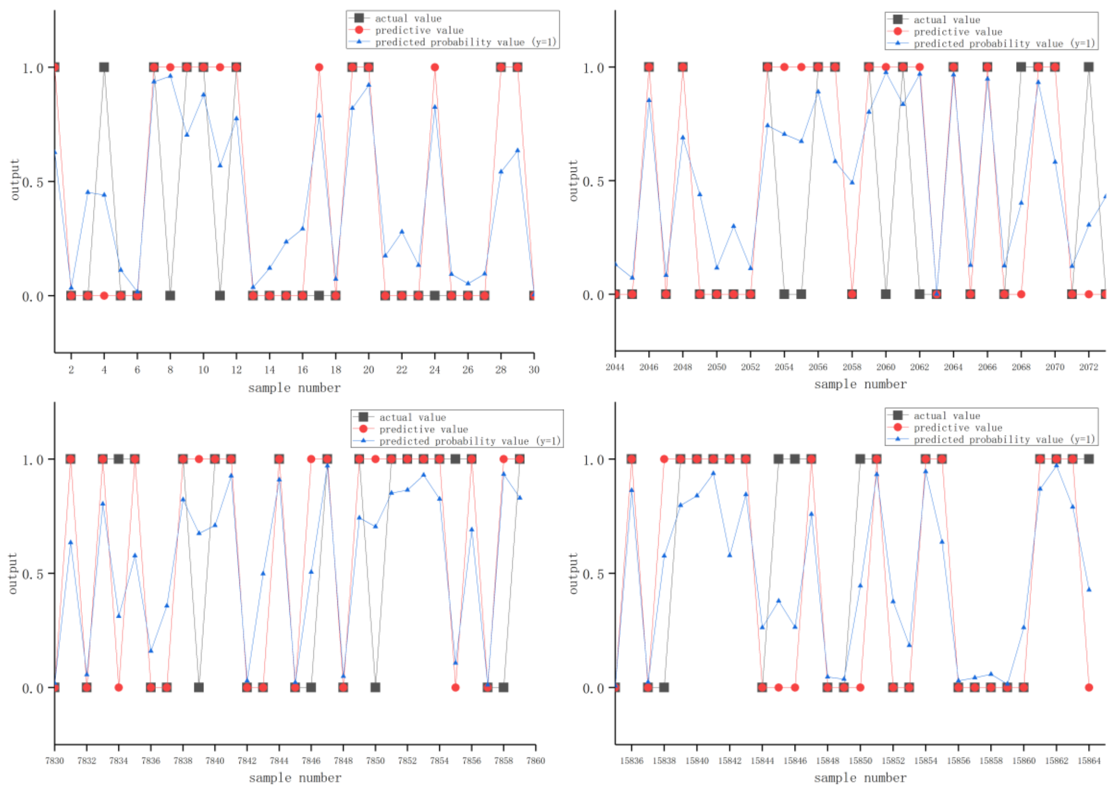


Fig. 4 Comparison charts of the predictive and actual values of the ANN

3.2.2 Radial Basis Function Neural Network

The input and output layer variables of the RBF neural network were the same as those of the ANN. After training, we obtained a hidden layer containing 10 units. The comparison charts of the predictive and actual values of the test set are shown in Figure 5.

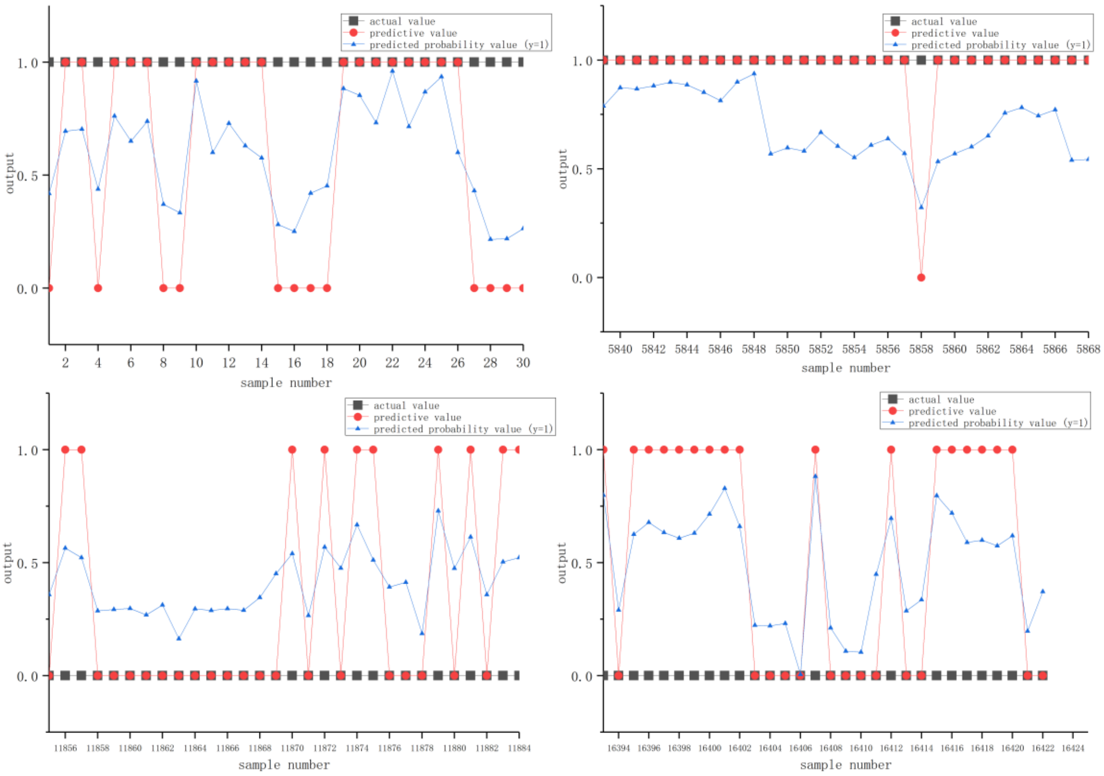
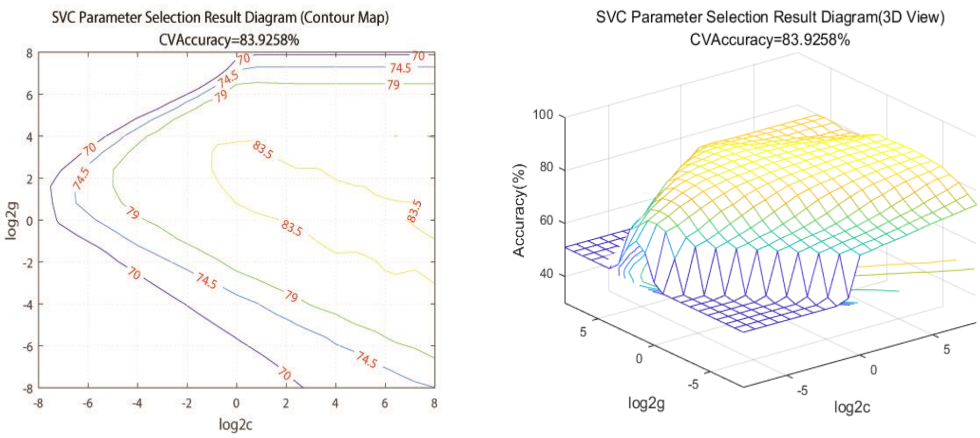


Fig. 5 Comparison charts of the predictive and actual values of the RBFN (part of the sample)

3.2.3 Support-Vector Machine

We used the LIBSVM package of the MATLAB software to construct the SVM. The model was constructed using the RBF kernel function for processing nonlinear data. We used the grid search method and 10-fold cross validation to select parameters and determine the penalty parameter *C* and the kernel parameter *g*. Figure 6 shows a contour map and a 3D view of the result of the SVC parameter selection. After calculation, the accuracy rate of the grid search method reached 83.9%, and the accuracy rate of cross-validation reached 82.6%.



(a) (b)

Fig 6. SVC parameter selection result: (a) contour map (b) 3D view

It can be seen from the results that the optimal values of *C* and *g* are 1.74 and 3.03, respectively. After setting the parameters to the optimal values, we performed SVM modeling and obtained the predicted values. Figure 7 shows the comparison charts of the actual and predicted values. After optimization, the total number of support vectors is 19,460, and the number of support vectors at the boundary is 17,260. After model training, the accuracy rate of the training set is 86.02%, and the accuracy rate of the test set is 84.27%, and the performance of the model is high.

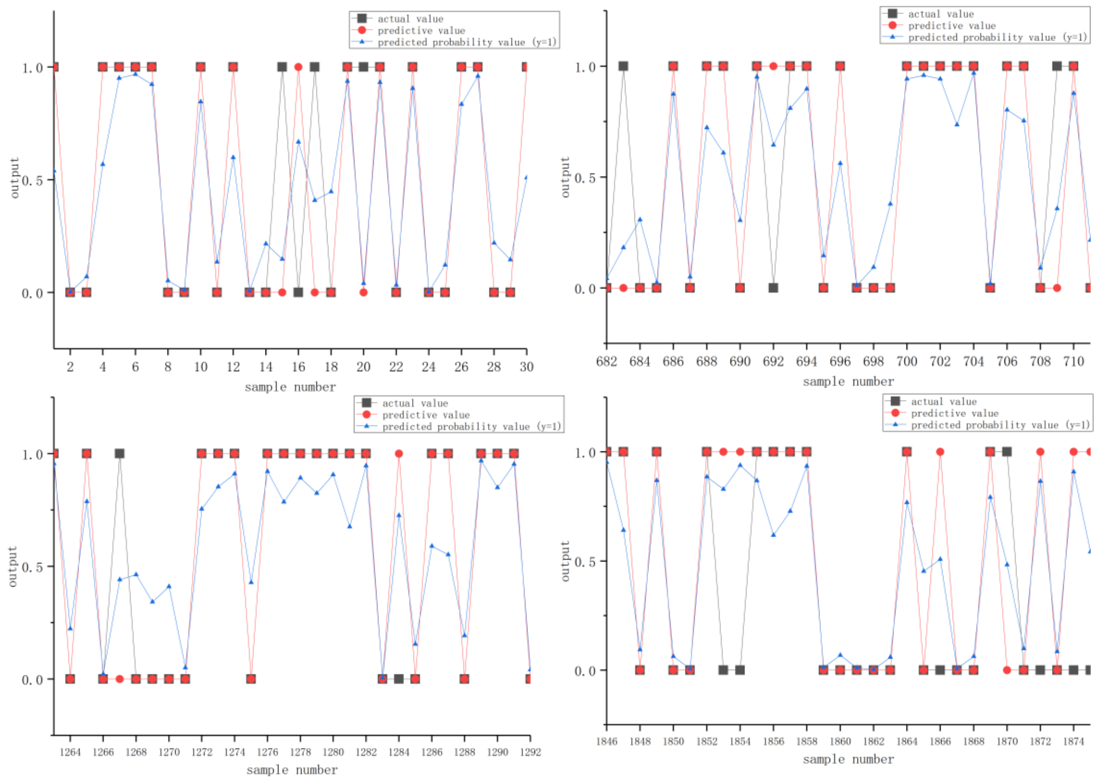


Fig. 7 Comparison charts of the predictive and actual values of the SVM (part of the sample)

3.2.4 Random Forest

We used the randomForest package in the R language to train random training samples. We then used cross-validation to determine the optimal parameters of the model and the number of optimal decision trees. Finally, we obtained the number of trees and the accuracy of the test and training data through cross-validation. As shown in Figure 8, when the number of decision trees is 400, the accuracy tends to be stable. We used the optimal number of decision trees to create the comparison charts of the actual and predicted values of the test set (Figure 9) and the average accuracy decline of 13 forest fire driving factors (Figure 10). It can be seen from Figure 10 that, among the main forest fire driving factors in China, the location variables that have the greatest influence on the occurrence of forest fires are longitude and latitude. Rainfall is the variable with the least influence on the occurrence of forest fires.

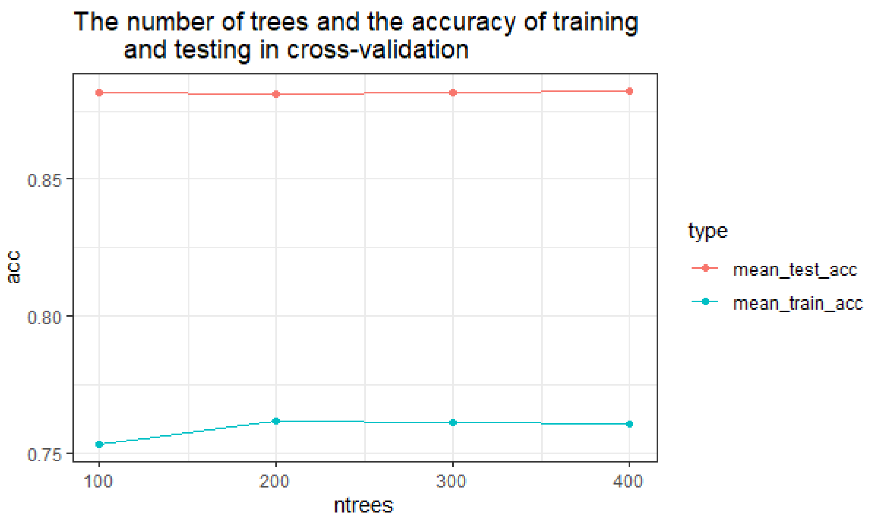


Fig. 8 The number of trees and the accuracy of training and testing in cross-validation

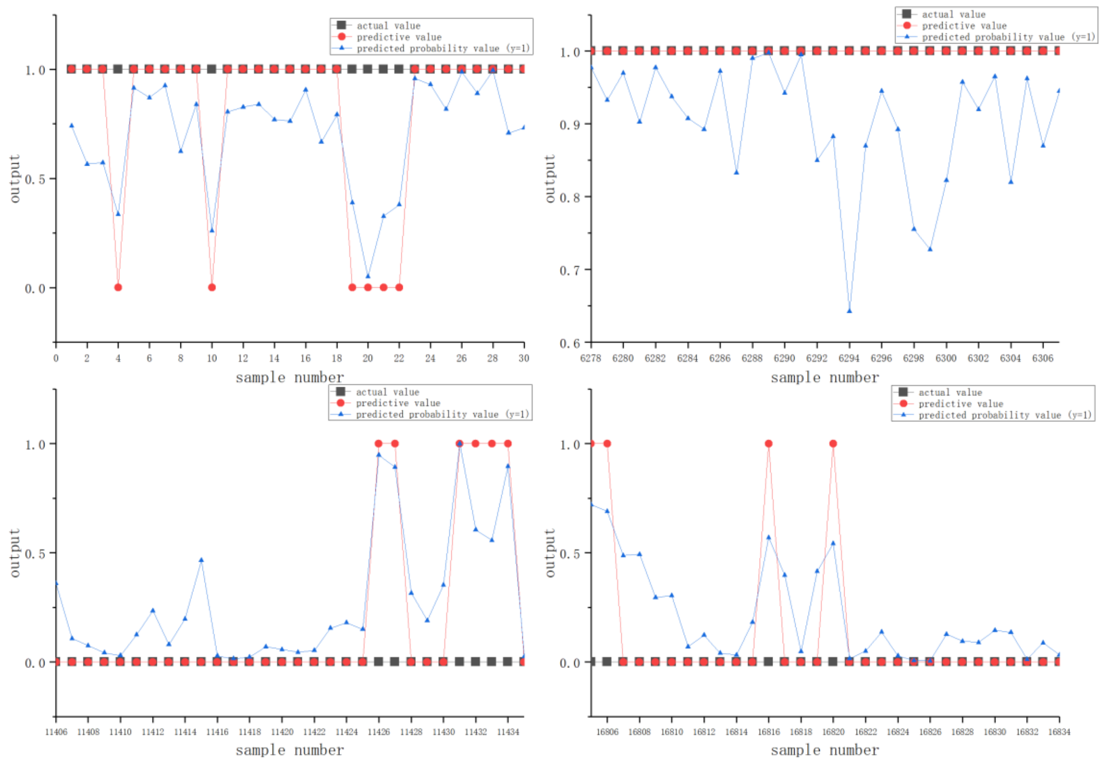


Fig. 9 Comparison charts of the predictive and actual values of the RF (part of the sample)

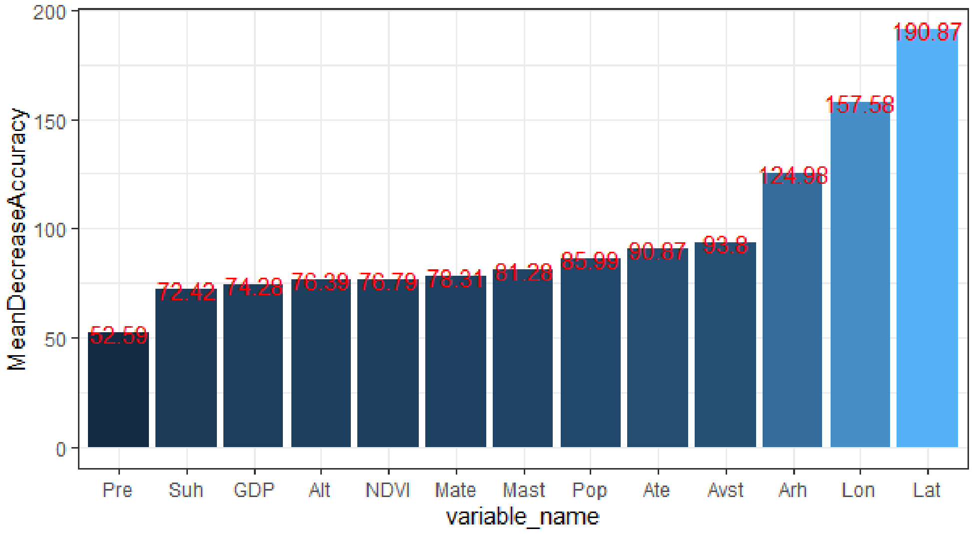


Fig. 10 Mean decrease accuracy of 13 variables

3.3 Accuracy Evaluation

We used the prediction results of the four models to construct a confusion matrix to obtain the accuracy, precision, recall, f1 value, and AUC value, as shown in Table 6. Figure 11 shows the visualization of the accuracy, precision, recall, and f1 values of the four models. Figure 12 shows the ROC curves of the four models. The accuracy and f1 value of each model are more than 75%, and the AUC value is more than 0.80. Thus, the performance of all four models is high. Among the four models, the RF model has the highest predictive ability, with an accuracy rate of 89.2%, an f1 value of 89%, and the highest AUC value, reaching 0.960. Compared with the other three models, the prediction ability of the RBF neural network is the lowest, with an accuracy rate of 75.8% and an AUC value of 0.840. As shown in Figures 11 and 12, the RF model outperforms the other three models. We therefore consider the RF model to be the most suitable of the four models for forest fire prediction in China.

Table 6: Evaluation results of the four models

Model	Accuracy (%)	Precision (%)	Recall (%)	f1 value (%)	AUC
ANN	83.0	85.4	79.6	82.4	0.904
RBFN	75.8	73.1	81.6	77.1	0.840
SVM	84.3	83.0	86.8	84.8	0.917
RF	89.2	90.2	87.9	89.0	0.960

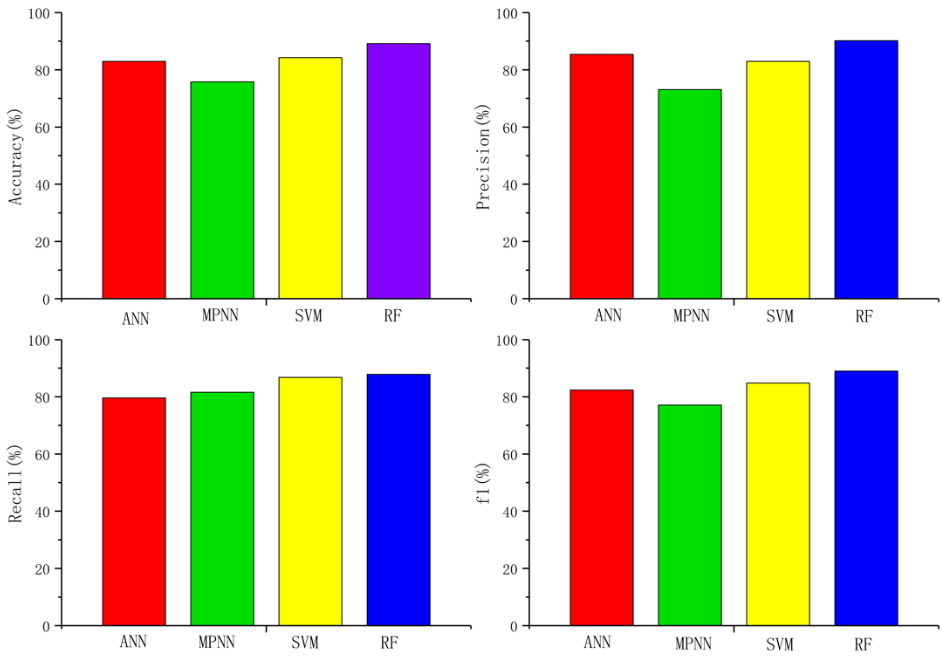


Fig 11. Comparison charts of accuracy, precision, recall, and f1 values of the four models

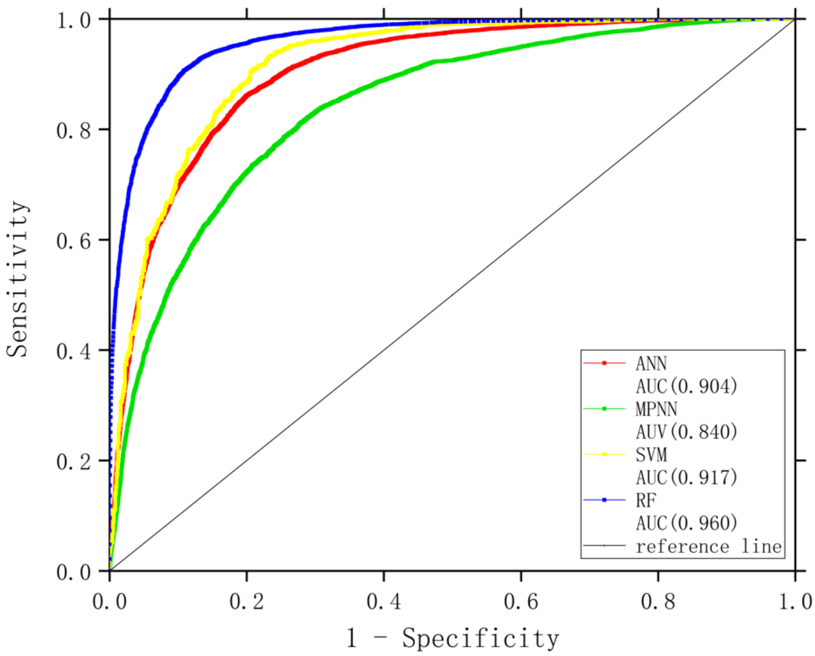


Fig. 12 ROC curves of the four models

3.4 Fire Risk Classification

After evaluating the accuracy of the four models, we used the RF model (highest accuracy) to obtain the probability of forest fire occurrence for the full sample. We used ArcGIS to draw a forest fire probability map (Figure 13) and a seasonal forest fire probability map (Figure 14) of China. Figure 13 shows that the high incidence of forest fires in China is mainly concentrated in the northeast (such as the Greater Xing'an Mountains region), the southeast (such as Guangdong, Jiangxi, and Fujian), and the southwest (such as Yunnan and Sichuan). On the whole, the probability of forest fires in eastern China is higher than that in western regions, and the probability of forest fires in the north and south is higher than that in central China. Figure 14 shows that the seasonal order of the probability of forest fires in China is, from highest to lowest, spring, winter, summer, and autumn. Spring and winter are the seasons with a high incidence of forest fires, which are mainly concentrated in northeast China (such as Heilongjiang Province) and southeastern China (such as Fujian Province and Guangdong Province).

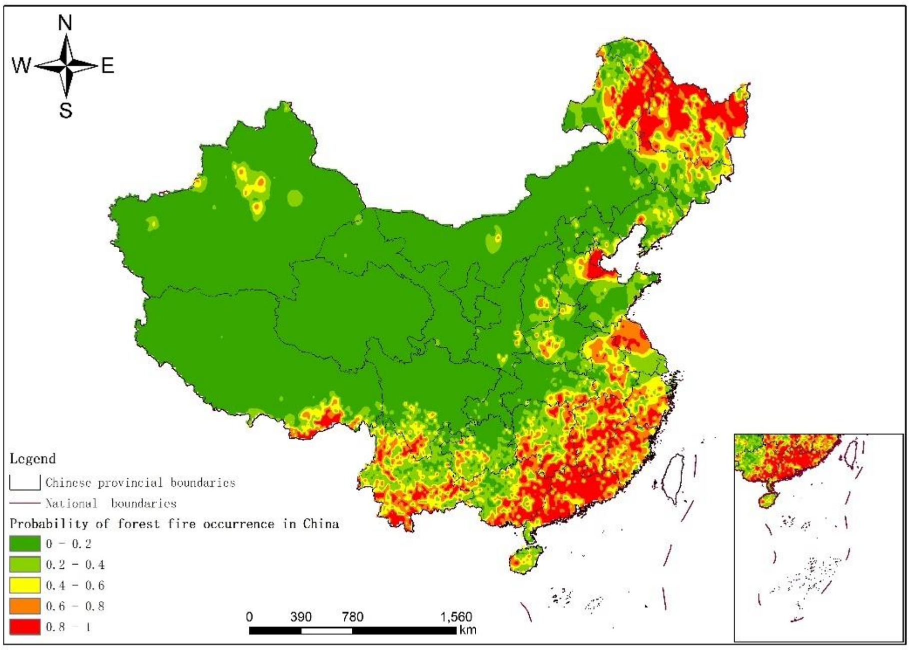


Fig. 13 Forest fire probability map of China based on the RF model

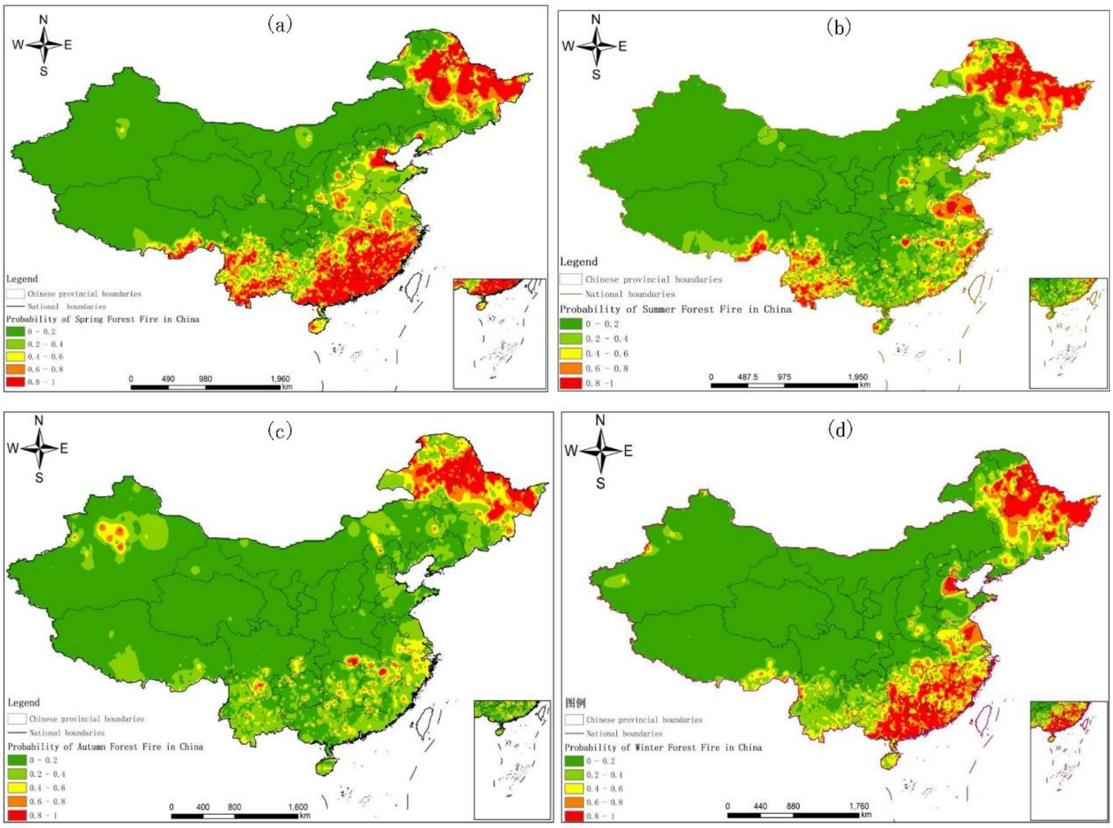


Fig. 14 Seasonal forest fire probability map of China based on the RF model: (a) spring (January, February, and March); (b) summer (April, May, and June); (c) autumn (July, August,

and September); and (d) winter (October, November, and December)

4. Discussion

4.1 Major Forest Fire Driving Factors in China and Their Impact

In this study, we selected over 20 factors affecting the occurrence of forest fires. These factors can be divided into six categories: location, meteorology, climate, terrain, society, and vegetation. Previous research has been conducted on the impact of these factors on forest fires [64-67]. We used the RF algorithm to filter the characteristics of these 20 forest fire driving factors and selected three or more variables in the five variable subsets as the main forest fire driving factors. In this experiment there are 13 main forest fire driving factors: longitude, latitude, average surface temperature, daily maximum surface temperature, cumulative precipitation, average relative humidity, sunshine hours, average temperature, daily maximum temperature, altitude, population, GDP, and NDVI.

The results show that the location factors (longitude and latitude) are the factors that have the greatest impact on the occurrence of forest fires. At present, there is little research that considers longitude and latitude as the main driving factors for forest fires. Some researchers, however, have confirmed that, as the latitude decreases, the number of forest fires increases [68-69]. China spans a large area, and the two factors of longitude and latitude reflect the regional differences in forest fires in China. Forest fires are more likely to occur at certain longitudes and latitudes.

Climate factors also have a great impact on forest fires, which is consistent with the findings of previous studies [70-73]. Temperature is one of the three necessary conditions for combustion. When the temperature reaches a certain level, forest fires are more likely to occur. The longer the sunshine hours, the higher the temperature, and the greater the probability of forest fires. Rainfall and average relative humidity are also among the main factors affecting forest fires [53][74-75].

Another type of driving factor is social and human factors (population and GDP). The larger the population, the greater the human activity in the region and the more likely it is that human factors cause forest fires. Catry et al. [76] (2007)and Sepulveda[77] (2001)reached the same conclusion. Altitude and vegetation (NDVI) also affect the occurrence of forest fires to varying degrees. Tian et al. [78] contend that forest fires mainly occur in low-altitude areas (2013). Chuvieco et al. [79] used NDVI as a driving factor for forest fires to estimate fuel moisture (2004). The greater the NDVI value, the higher the vegetation coverage; and the greater the flammability of the tree species, the more likely such trees are to cause problems related to forest fires.

In this study, we used a RF algorithm for feature selection and eliminated seven variables: average wind speed, maximum wind speed, aspect, slope, distance from fire point to highway, closest distance from fire point to residential area, and special festivals. In this experiment, these factors have little effect on the occurrence of forest fires. It may be that these factors change with time and space, however. Other studies have found that these factors are among the main drivers of forest fires[80][81]. We believe that this may be due to the difference in the selected data and the difference in the method of feature selection. In future research, a variety of feature-screening methods and analysis of different regions may be used to obtain more comprehensive results.

4.2 Optimal Choice of Forest Fire Prediction Model

We entered the forest fire driving factors selected by feature selection into the four models (ANN, RBF neural network, SVM, and RF) for training. We then evaluated them using five criteria: accuracy, precision, recall rate, f1 value, and AUC value. We selected the RF model as the optimal choice for forest fire prediction. The accuracies of all four models are above 75%, which means that they are all reliable. The RF model, however, exhibits the greatest prediction ability. The RBF neural network model has the lowest prediction performance.

Samaher et al. [25] used a cascade correlation network, multilayer perceptron neural network, polynomial neural network, RBF, and SVM for forest fire prediction(2018). They found the prediction performance of the SVM was the highest, and the performance of RBF was the lowest, which is consistent with our conclusion. Sakr et al. [36] used an SVM and an ANN to predict the fire risk in Lebanon (2011). Their results showed that the performance of the SVM model was higher than that of the ANN model. This finding is similar to ours. Paulo et al. [8] used RF and SVM models to predict forest fires (2007). They concluded that the performance of the SVM model was higher than that of the RF. Their finding is different from our results. A possible reason for this difference is that Paulo et al. chose four types of weather factors and made predictions about small areas, whereas we chose 13 types of factors and made predictions about a large area. The choice of variables and the difference in the sample size affect the model training.

Bisquert et al. [82] used an ANN to establish a forest fire hazard model with a highest accuracy rate of 76%, which is lower than the accuracy of our model (83%) (2012). Hong [83] used an SVM algorithm to analyze Dayu County in southwestern Jiangxi Province, China (2018). The results show that the AUC value of the SVM is 0.75, which is lower than the value in our model (0.92). Pourtaghid et al.[45] used an RF to create forest fire sensitivity analysis with a prediction accuracy of 72.8% (2016). Our model reached a prediction accuracy of 89.2%.

The four models we selected all exhibit high predictive capabilities. The main reason may be that appropriate multi-dimensional variables has been screened out and the data sample size is large, which makes the training of each model more accurate and reliable.

There are also differences in the characteristics of these four models. The ANN and RBF neural network models can be trained very quickly, and they can handle samples with a large amount of data, but their accuracy in this experiment is relatively low. In subsequent studies, particle swarms or genetic algorithms could be used to improve the accuracy of these models. The SVM model has a high predictive ability, but it also has shortcomings. The higher the model complexity, the lower the calculation speed. It takes a longer time in this model to obtain the optimal parameters when processing large sample data. We will consider using other algorithms to optimize SVM in the future. The RF model exhibited excellent expressive power in this experiment. It can quickly process large data samples while ensuring high prediction accuracy.

4.3 Recommendations for Forest Fire Prevention

We produced a probability map of forest fires in China that shows that the highest incidences of forest fires are in the northeast (Heilongjiang Province and the northern Inner Mongolia Autonomous Region), the southeast (Fujian Province, Guangdong Province, and Jiangxi Province), and Yunnan Province. The pattern of forest fire points presents a spatial clustering distribution. Ma et al. obtained similar results [4]. For these high-incidence areas, watch towers and monitoring equipment should be added for monitoring and management. Moreover, the length of the forest fire barrier net should be increased to reduce the spread of fires. In addition, the number of fire brigades and fire vehicles should be increased to enhance the disaster-mitigation capabilities. Regarding seasonal risks of forest fires, forest fire prevention and control should be emphasized in spring and winter. Strengthening fire-prevention management during these periods would mainly involve strengthening the management of human activities to reduce human-made forest fires and improving publicity and education, such as the addition of fire-prevention signs.

This study has some shortcomings, and there is scope for improvement. One of the three elements of fire is combustible fuel. For the selection of forest fire driving factors, however, there is currently no way to obtain data on fuel load and other related factors. Thus, this experiment lacked relevant data such as combustible load, particle size of combustible material, and combustible tree species. If possible, in future research, such data could be added to the forest fire prediction model.

This study selected four kinds of machine-learning algorithms for the forest fire prediction model. Other applicable machine-learning algorithms could be used in future experiments. In addition, the ability of these machine-learning algorithms to analyze spatial heterogeneity is relatively weak. Subsequent research could use geographically weighted regression to build a high-precision forest fire prediction model.

5. Conclusion

This study determined the main driving factors of forest fire occurrence in China through feature selection. The main factors that affect the occurrence of forest fires to varying degrees are meteorological, topographic, man-made, and vegetation factors. We built four forest fire prediction models using the following machine-learning algorithms: an artificial neural network, a radial basis function neural network, a support-vector machine, and a random forest. The results of the evaluation show that the accuracy of all the models is higher than 75%. These models can be used to build forest fire prediction models. Among the four models, the RF model has the highest comprehensive predictive ability, with an accuracy of 89.25%. It is therefore the optimal choice for a forest fire prediction model in China.

We used the RF model to predict the probabilities of forest fires in China. Based on these probabilities, we drew a map of the probability of forest fire occurrence in China and a map of the probability of forest fires in China by season (spring, summer, autumn, and winter). Finally, based on these maps, we identified the high-incidence areas and areas at risk of forest fires. We then put forward fire-prevention recommendations for the corresponding regions and seasons.

This research helps in understanding the main forest fire driving factors in China. It provides a reference for the selection of high-precision forest fire prediction models. In addition, it provides suggestions on the time and location of forest fire prevention in China. Moreover, this study provides guidance for China's forest fire prevention and control work.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Funding Statement

This research was jointly supported by the medium long-term project of “Precision Forestry Key Technology and Equipment Research” (No. 2015ZCQ-LX-01) and the National Natural Science Foundation of China (No. U1710123).

Authors' contributions

Yudong Li performed the experiment and wrote the manuscript;

Zhongke Feng contributed to the conception of the study;

Ziyu Zhao and Shilin Chen helped perform the analysis with constructive discussions;

Hanyue Zhang helped perform the data analysis.

Acknowledgments

We would like to acknowledge support from Beijing Key Laboratory for Precision Forestry, Beijing Forestry University, as well as all the people who have contributed to this paper.

References

[1] Bhusal, Satish & Mandal, Ram. (2020) Forest fire occurrence, distribution and future risks in Arghakhanchi district, Nepal. Journal of Geography. 2(1):10-20.

[2] B. K. Singh, N. Kumar and P. Tiwari. (2019) "Extreme Learning Machine Approach for Prediction of Forest Fires using Topographical and Metrological Data of Vietnam," 2019 Women Institute of Technology Conference on Electrical and Computer Engineering (WITCON ECE), Dehradun Uttarakhand, India,. pp. 104-112. doi: 10.1109/WITCONECE48374.2019.9092926.

[3] D Mandallaz, R Ye. (1997) Prediction of forest fires with Poisson models, Canadian Journal of

Forest Research. 27(10): 1685-1694.https://doi.org/10.1139/x97-103.

[4] Ma, W., Feng, Z., Cheng, Z., Chen, S., & Wang, F. (2020) Identifying Forest Fire Driving Factors and Related Impacts in China Using Random Forest Algorithm. Forests. 11(5), 507. doi:10.3390/f11050507.

[5] Dimopoulou, M.; Giannikos, I. (2004) Towards an integrated framework for forest fire control. Eur. J. Oper. Res. 152: 476–486.

[6] Flannigan, M.D.; Krawchuk, M.A.; Groot, W.J.D.; Wotton, B.M.; Gowman, L.M. (2009) Implications of changing climate for global wildland fire. Int. J. Wildland Fire. 18:483–507.

[7] Z. Q. Xu, X. Y. Su, Y. Zhang. (2012) Forest Fire Prediction Based on Support Vector Machine. Chinese Agricultural Science Bulletin.28(13):126-131.

[8] Paulo Cortez, Anibal Morais. (2007) A Data Mining Approach to Predict Forest Fires using Meteorological Data. New Trends in Artificial Intelligence. 512-523.

[9] Barik A, Baidya Roy S. (2020) Effects of meteorology on forest fires in India: A modeling study[C]//EGU General Assembly Conference Abstracts. 18317.

[10] Chetehouna K. (2020) Overview of The Forest Fire Research[C]//Proceeding International Conference on Science and Engineering. 3: xvii-xvii.

[11] Shang C, Wulder M A, Coops N C, et al. (2020) Spatially-Explicit Prediction of Wildfire Burn Probability Using Remotely-Sensed and Ancillary Data. Canadian Journal of Remote Sensing. 1-17.

[12] Z.W. SU, A.Q. LIU et al. (2016) Driving factors and spatial distribution patteren of forest fire in Fujian Province. JOURNAL OF NATURAL DISASTERS. 25(2):110-119.

[13] M. Yu, (2016) The establishment and study of Regional forest fire prediction and forecasting

model. Beijing, Beijing Forestry University.

[14] Y. Song. (2018) Research on Forest Fire Drivers and Models in Heilongjiang Province. Herbing, Northeast Forestry University.

[15] CV Garcia, PM Woodard, SJ Titus, WL Adamowicz and BS Lee. (1995) A Logit Model for Predicting the Daily Occurrence of Human Caused Forest-Fires. International Journal of Wildland Fire. 5(2):101-111.

[16] H. Peng , M.C. Shi , Y. Sun, Mike Wotton. (2014) Lightning Fire Forecasting Model of Daxing'an Mountain Based on Logistic Model. Journal of Northeast Forestry University. 42(7):166-169.

[17] W.G. Wang, J.H. Pan, Y.Y Feng, Z. Li, L. L. Dong. (2017) Model and Zoning of Fire Risk in Gansu Province based on GWLR and MODIS Imagery. Remote Sensing Technology and Application. 32(03):514-523.

[18] Futao Guo, Selvara Selvalakshmi, Fangfang Lin, Guangyu Wang, Wenhui Wang, Zhangwen Su, Aiqin Liua.(2016)Geospatial information on geographical and human factors improved anthropogenic fire occurrence modeling in the Chinese boreal forest. Canadian Journal of Forest Research. 46(4): 582-594. https://doi.org/10.1139/cjfr-2015-0373.

[19] K.Z. Liu, L.F. Shu, F.J. Zhao et al. (2017) Research on spatial distribution of forest fire based on satellite hotspots data and forecasting model. Journal of Forestry Engineering, 2:128-133.

[20] B.Q. LIAO, J. WEI et al. (2008) Logistic and ZIP Regression Model for Forest Fire Data. FIRE SAFETY SCIENCE.3:143-149.

[21] D Mandallaz, R Ye. (1997) Prediction of forest fires with Poisson models. Canadian Journal of Forest Research.27(10): 1685-1694. https://doi.org/10.1139/x97-103.

[22] F.T. GUO, H.Q. HU et al. (2010) Relationship between forest lighting fire occurrence and weather

factors in Daxing’an Mountains based on negative binomial model and zero-inflated negative binomial models. Chinese Journal of Plant Ecology.21(01):159-164.

[23] H.L. Liang ,Y.R. Lin, G. Yang et al. (2016) Application of Random Forest Algorithm on the Forest Fire Prediction in Tahe Area Based on Meteorological Factors. SCIENTIA SILVAE SINICAE.52(01):89-98.

[24] Z.Q. Xu. (2012) Forest Fire Area Prediction Based on Support Vector Machine. Beijing, Beijing Forestry University.

[25] Samaher Al_Janabia et al. (2018) Assessing the suitability of soft computing approaches for forest fires prediction. Applied Computing and Informatics. (14): 214-224.

[26] Volkan Sevinca, Omer Kucukb, Merih Goltasc. (2020) A Bayesian network model for prediction and analysis of possible forest fire causes. Forest Ecology and Management. (457):1-11.

[27] T. Artés, A. Cencerrado, A. Cortés, T. Margalef. (2016) Time aware genetic algorithm for forest fire propagation prediction: exploiting multi-core platforms, Concurrency and Computation: Practice and Experience.

[28] D.T. Bui, Q.T. Bui, Q.P. Nguyen, B. Pradhan, H. Nampak, P.T. Trinh, (2017) A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area, Agric. For. Meteorol. 233: 32-44

[29] M. Denham, A. Cortés, T. Margalef, E. Luque, (2008) Applying a dynamic data driven genetic algorithm to improve forest fire spread prediction. International Conference on Computational Science. Springer Berlin Heidelberg. pp.36-45.

[30] H. Hong, S.A. Naghibi, M.M. Dashtpagerdi, H.R. Pourghasemi, W. Chen, (2017) A comparative

assessment between linear and quadratic discriminant analyses (LDA-QDA) with frequency ratio and weights-of-evidence models for forest fire susceptibility mapping in China. Arab. J. Geosci.10 (7): 167.

[31] A.M. Özbayog ˇlu, R. Bozer, (2012) Estimation of the burned area in forest fires using computational intelligence techniques, Proc. Comput. Sci. 12 :282–287.

[32] H. Soliman, K. Sudan and A. Mishra. (2010) A smart forest-fire early detection sensory system: Another approach of utilizing wireless sensor and neural networks. SENSORS, 2010 IEEE, Kona, HI. pp. 1900-1904.doi: 10.1109/ICSENS.2010.5690033.

[33] Çetin Elmas, Yusuf Sönmez. (2011) A data fusion framework with novel hybrid algorithm for multi-agent Decision Support System for Forest Fir. Expert Systems with Applications.38(8):9225-9236. https://doi.org/10.1016/j.eswa.2011.01.125.

[34] Onur Satir, Suha Berberoglu & Cenk Donmez (2016) Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem, Geomatics. Natural Hazards and Risk. 7: 1645-1658, doi: 10.1080/19475705.2015.1084541.

[35] Maeda, E. E., Formaggio, A. R., Shimabukuro, Y. E., Arcoverde, G. F. B., & Hansen, M. C. (2009). Predicting forest fire in the Brazilian Amazon using MODIS imagery and artificial neural networks. International Journal of Applied Earth Observation and Geoinformation, 11(4), 265–272. doi: 10.1016/j.jag.2009.03.003.

[36] Sakr G E, Elhajj I H, Mitri G. (2011) Efficient forest fire occurrence prediction for developing countries using two weather parameters. Engineering Applications of ArtificialIntelligence.24(5): 888-894.

[37] B.C. Ko, K.H. Cheong, J.Y. Nam. (2009) Fire detection based on vision sensor and support vector

machines. Fire Saf. J. 44 (3): 322-329.

[38] G.E. Sakr, I.H. Elhajj, G. Mitri, Efficient forest fire occurrence prediction for developing countries using two weather parameters. Eng. Appl. Artif. Intell. 24.

[39] D.W. Xie, S.L. Shi,(2014)Prediction for burned area of forest fires based on SVM model, in: Applied Mechanics and Materials. Trans Tech Publications. 513(5):4084-4089.

[40] J. Zhao, Z. Zhang, S. Han, C. Qu, Z. Yuan, D. Zhang. (2011) SVM based forest fire detection using static and dynamic features. Computer Sci. Inform. Syst. 8(3): 821–841.

[41] Cutler DR , et al. (2007) Random forests for classification in Ecology. Ecology.88(11):2783-2792.

[42] Prasad AM, et al. (2006) Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems .9(2):181-199.

[43] Rodrigucs M, Dc la Riva J. (2014) An insight into machines learning algorithms to model humarrcaused wildfire or currence. Environmental Modelling & Software,57:192-201.

[44] LIANG H L, LIN Y R, YANG G, et al.(2016) Application of random forest algorithm on the forest fire prediction in Tahe area based on meteorological factors. Scientla Silvae Sinicae.52(1):89-98.

[45] Pourtaghi, Z. S., Pourghasemi, H. R., Aretano, R., & Semeraro, T. (2016) Investigation of general indicators influencing on forest fire and its susceptibility modeling using different data mining techniques. Ecological Indicators. 64:72–84. doi:10.1016/j.ecolind.2015.12.030.

[46] Tian, X., Zhao, F., Shu. L. (2013) Wang, M. Distribution characteristics and the influence factors of forest fires in China. For. Ecol. Manag. 310:460–467.

[47] Chang, Y., Zhu, Z., Bu, R., Li, Y., Hu, Y. (2015) Environmental controls on the characteristics of mean number of forest fires and mean forest area burned (1987–2007) in China. For. Ecol. Manag.

356:13–21.

[48] Zhong. M., Fan, W., Liu. T., Li. P. (2003) Statistical analysis on current status of China forest fire safety. Fire Saf. J. 38:257–269.

[49] Aifeng, L.U. (2011) Study on the relationship among forest fire, temperature and precipitation and its spatial-temporal variability in China. Agric. Sci. Technol. 12:1396–1400.

[50] Ying. L., Han. J., Du. Y., Shen, Z. (2018) Forest fire characteristics in China: Spatial patterns and determinants with thresholds. For. Ecol. Manag.424:345–354.

[51] X.K. Ren, Chinese Geography, Chinese Press, ISBN:9787507528800.

[52] F.T. Guo, Z.W. Su, G.Y. Wang et al. (2017) Understanding fire drivers and relative impacts in different Chinese forest ecosystems. Science of the Total Environment.605: 411-425.

[53] GUO F, WANG G, SU Z, et al. (2016) What drives forest fire in Fujian, China? Evidence from logistic regression and random forests. International Journal of Wildland Fire.25(5):505-519.

[54] Chang Y, Zhu Z L, Bu R C, et al. (2013) Predicting fire occurrence pat-terns with logistic regression in Heilongjiang Province, China. Landscape Ecology.28(10) : 1989 -2004．

[55] XU X L. (2018) Spatial distribution data set of quarterly vegetation index (NDVI) in China. Data Registration and Publishing System of Resources and Environmental Science Data Center of Chinese Academy of Sciences (http://www.resdc.cn/DOI).

[56] W.Y MA, Z.K. FENG et al. (2020) Study on driving factors and distribution pattern of forest fires in Shanxi province. Journal of Central South University of Forestry & Technology.1-13.

[57] ERIDUNTONGLAGA. (2013) Analysis on landscape pattern of woodland based on DEM in the Daqingshan Mountains Inner Mongolia. Huhhot: Inner Mongolia Normal University.

[58] A. Subasi, (2007) EEG signal classification using wavelet feature extraction and a mixture of

expert model. Expert Syst. Appl. 32 (4): 1084-1093.

[59] D. Chen. (2019) Prediction of Forest Fire Occurrence in Daxing'an Mountains Based on Logistic Regression Model," FOREST RESOURCES MANAGEMENT. (01):116-122.

[60] Q.Q. Wang, J.T. Tong, L. Zhang et al. (2020) Seismic data denoising using multi-layer perceptron. Oil Geophysical Prospecting.55(02):272-281+228.

[61] D.S. Broomhead, D. Lowe, (1988) Radial basis functions, multi-variable functional interpolation and adaptive networks (No. RSRE-MEMO-4148). Royal signals and radar establishment Malvern (United Kingdom).

[62] X.C. Wang et al. (2013) 43 Cases of MATLAB neural network analysis. Beijing, Bei hang University Press.

[63] Yudong Li, Zhongke Feng , Shilin Chen, Ziyu Zhao, and Fengge Wang. (2020) Application of the Artificial Neural Network and Support Vector Machines in Forest Fire Prediction in the Guangxi Autonomous Region, China. Discrete Dynamics in Nature and Society. https://doi.org/10.1155/2020/5612650.

[64] Ganteaume A, Camia A, Jappiot M, et al. (2013) A review of the main driving factors of forest fire ignition over Europe. Environmental Management. 51(3): 651-662.

[65] Syphard AD, Radeloff VC, Keuler NS, Taylor RS, Hawbaker TJ, Stewart SI, Clayton MK (2008) Predicting spatial patterns of fire on a southern California landscape. Int J Wildland Fire 17:602–613.

[66] Pew KL, Larsen CPS (2001) GIS analysis of spatial and temporal patterns of human-caused wildfires in the temperate rainforest of Vancouver Island, Canada. Forest Ecol Manag 140:1-18.

[67] Dickson BG, Prather JW, Xu Y, Hampton HM, Aumack EN, Sisk TD (2006) Mapping the

probability of large fire occurrence in northern Arizona. USA Landsc Ecol 2:747-761.

[68] Ying. L., Han. J.; Du. Y., Shen, Z. (2018) Forest fire characteristics in China: Spatial patterns and determinants with thresholds. For. Ecol. Manag. 424:345–354.

[69] Prasad, A.M.; Iverson, L.R.; Liaw, A. (2006) Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems. 9:181-199.

[70] Guo F, Su Z, Wang G, et al. (2016) Wildfire ignition in the forests of southeast China: Identifying drivers and spatial distribution to predict wildfire likelihood. Applied Geography. 12-21.

[71] Liu, Z., Yang, J., Chang, Y., Weisberg, P. J., & He, H. S. (2012). Spatial patterns and drivers of fire occurrence and its future trend under climate change in a boreal forest of northeast China. Global Change Biology, 18:2041-2056.

[72] Syphard, A. D., Radeloff, V. C., Keuler, N. S., Taylor, R. S., Hawbaker, T. J., Stewart, S. I., et al. (2008) Predicting spatial patterns of fire on a southern California landscape. International Journal of Wildland Fire.17:602-613.

[73] Ying L, Han J, Du Y, et al. (2018) Forest fire characteristics in China: Spatial patterns and determinants with thresholds. Forest Ecology and Management.345-354.

[74] Zumbrunnen, T., Pezzatti, G. B., Men?endezd, P., Bugmann, H., Bürgi, M., & Conedera, M. (2011) Weather and human impacts on forest fires: 100 years of fire history in two climatic regions of Switzerland. Forest Ecology and Management.261:2188-2199.

[75] Cardille J A, Ventura S J, Turner M G, et al. (2001) ENVIRONMENTAL AND SOCIAL FACTORS INFLUENCING WILDFIRES IN THE UPPER MIDWEST, UNITED STATES. Ecological Applications.11(1): 111-127.

[76] Catry F X, Damasceno P, Silva J S. (2007) Spatial distribution patterns of wildfire ignitions in

Portugal. Available at: http://www. eufirelab.org/toolbox2/library/upload/2380.pdf, 2009-08.

[77] Sepulveda B J I, Meza S R, Zuniga C W R et al. (2001) GIS to Determine Risk of Forest Fires in Northwestern Mexico. Mexico: Technical Publication. 37.

[78] Tian, X.; Zhao, F.; Shu, L.; Wang, M. (2013) Distribution characteristics and the influence factors of forest fires in China. For. Ecol. Manag. 310: 460-467.

[79] Chuvieco, E.; Cocero, D.; Riaño, D.; Martin, P .; Martínez-Vega, J.; de la Riva, J.; Pérez, F. (2004) Combining ndvi and surface temperature for the estimation of live fuel moisture content in forest fire danger rating. Remote Sens. Environ. 92: 322-331.

[80] Maingi K J, Henry M C. (2007) Factors influencing wildfire occurrence and distribution in eastern Kentucky, USA. International Journal of Wildland Fire, 16(1): 23-33. doi: 10.1071/ WF06007.

[81] Avilaflores D Y, Pompagarcia M, Antonionemiga X, et al. (2010) Driving factors for forest fire occurrence in Durango State of Mexico: A geospatial perspective. Chinese Geographical Science. 20(6): 491-497.

[82] Bisquert M, Caselles E, Sanchez J M, et al. (2012) Application of artificial neural networks and logistic regression to the prediction of forest fire danger in Galicia using MODIS data. International Journal of Wildland Fire. 21(8): 1025-1029.

[83] Hong, H., et al., (2018) Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models. The case of Dayu County, China. Science of The Total Environment. 630:1044-1056.